Automatic Singer Identification Based on Speech-Derived Models

Wei-Ho Tsai and Hsin-Chieh Lee

***Abstract*—Automatic singer identification (SNID) aims to determine who among a set of singers perform a given music recording. So far, most existing SNID methods follow a framework stemming from speaker identification (SPID) research, which models each person's characteristics using his/her voice data. This framework, however, is impractical in many SNID applications, because acquiring solo a cappella from each singer is usually not as feasible as collecting spoken data from each speaker in SPID applications. In view of the easy availability of spoken data, this work investigates the possibility of modeling singers' voices using spoken data instead of singing data. However, our experiment found it difficult to replace singing data fully by using spoken data in singer voice modeling, due to the significant difference between singing and speaking for most people. Thus, we propose an alternative solution based on the use of few available singing data. The idea is to modify speech-derived voice models using MAP adaptation on few singing data, so that the adapted voice models can cover performers' singing characteristics. Our experiments found that most of the singing clips can be correctly identified using the adapted voice models.**

***Index Terms*—Model adaptation, singer identification, speaker identification.**

I. Introduction

Human voice recognition has long been an important research topic. With the increasing demand of audio information retrieval, this research topic has recently been extended from speaker identification (SPID) [1] to singer identification (SNID) [2], [3]. Currently, most existing SNID methods [2], [3] follow a framework stemming from SPID research, which models each person's characteristics using his/her, voice data. This framework, however, is impractical in many SNID applications, because acquiring solo a cappella from each singer is usually not as feasible as collecting spoken data from each speaker in SPID applications.

Imagine that when your family or friends gather to sing at a Karaoke, you may like to record everyone's performance into CDs or DVDs to capture memories of the pleasant time. For the audio in CDs or DVDs to be searchable, audio data would preferably be written in separate tracks, each labelled with the respective singer(s). As manually labelling a long audio stream is time-consuming, you would like to have an automated system for identifying singers in the recordings. However, existing SNID systems require that sufficient singing voice data of each singer be collected in advance to perform the so-called training. If no or insufficient singing data are collected from your family and friends beforehand, it becomes infeasible to apply SNID in this case. Instead of relying on pre-collected singing data, this study investigates an alternative solution to the above scenario. The idea is that if there are available speech data from each singer (which may be easy to acquire), can we design an SNID system that characterizes singers' voices using the speech data? Such an idea can also be used in identifying popular singers. For most popular music, it is almost impossible to acquire singers' solo voices, since audio data contain background accompaniments during vocal passages. However, there are many opportunities to acquire their spoken data, such as, from TV interviews or press conferences. Thus, singers' spoken data may be used to train an SNID system.

To the best of our knowledge, there is no prior literature devoting to this problem. Most related work [4,5] investigate the differences between singing and speech. Some studies develop methods for singing voice synthesis [6,7], and some discuss how to convert speech into singing [8], according to the specified melody. This work investigates two SNID systems trained using spoken data. However, as a first step toward our research objective, we do not deal with the interference of background accompaniment for SNID, and consider a cappella music only.

The rest of this paper is organized as follows. Section 2 introduces an intuitive SNID system based on speech-derived singer models. Section 3 describes the proposed SNID system using spoken data as well as few singing data for singer voice modelling. Then, Section 4 discusses our experiment results. In Section 5, we present our concluding remarks.

II. An Intuitive SNID System

Following the most popular SPID framework [9], we can build an intuitive SNID system as shown in Fig. 1. The system operates in two phases: training and testing. During training, a group of N persons is represented by Gaussian mixture models (GMMs), λ1, λ2, λN. It is known that GMMs provide good approximations of arbitrarily shaped densities of spectrum over a long span of time [9], and hence can reflect the vocal tract configurations of individual persons. The parameters of each GMM, say λi, are estimated using the speech utterances of the i-th person. The estimation consists of k-means initialization and Expectation-Maximization (EM) [10].

Prior to Gaussian mixture modeling, audio waveforms are

Manuscript received April 13, 2012; revised May 23, 2012

W. H. Tsai and H. C. Lee are with the Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan, (Tel.: +886-2-27712171 x 2257; fax: +886-2-27317120; e-mail: whtsai@ntut.edu.tw)

converted, frame-by-frame, into Mel-scale frequency cepstral coefficients (MFCCs). Since MFCCs carry less information on pitch than vocal tract configuration, they should be able to absorb the discrepancy between singing and speech in the pitch variations. Given a test singing recording represented by MFCCs X, the system decides in favor of singer I* when the maximum likelihood (ML) condition in Eq. (1) is satisfied:

$$I^{*} = \arg\max_{1\le i\le N} \log P(\mathbf{X} \mid \lambda_i) \qquad (1)$$

where P(X|λi) represents the probability that X is generated by λi.

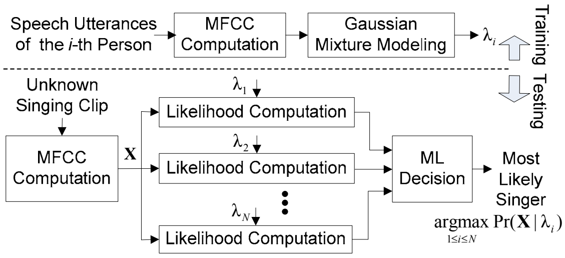

Fig. 1. An intuitive SNID system, which characterizes singers' voices using their speech utterances.

III. An Improved SNID System Based on Model Adaptation

Our experiments, detailed in Sec. 4, find that the intuitive system performs rather poor, since a person's singing voice can be significantly different from his/her speech voice. To improve the system, we propose adapting each person's GMM by using a few of his/her singing voice data. The adaptation is based on the Maximum A Posterior (MAP) estimation of GMM parameters [11]. Since the amount of available singing data for adaptation is assumed to be very limited, we only adapt the mean vectors of GMMs. For the *i*-th person's GMM, the mean vector of the *k*-th mixture is updated using

$$\hat{\boldsymbol{\mu}}_i^{(k)} = \frac{\tau_i^{(k)}}{\tau_i^{(k)}+\gamma}\,\bar{\boldsymbol{\mu}}_i^{(k)} + \frac{\gamma}{\tau_i^{(k)}+\gamma}\,\boldsymbol{\mu}_i^{(k)} \qquad (2)$$

$$\tau_i^{(k)} = \sum_{\ell=1}^{L} \Pr(k \mid \mathbf{x}_\ell, \lambda_i) \qquad (3)$$

$$\bar{\boldsymbol{\mu}}_i^{(k)} = \frac{1}{\tau_i^{(k)}} \sum_{\ell=1}^{L} \Pr(k \mid \mathbf{x}_\ell, \lambda_i)\,\mathbf{x}_\ell \qquad (4)$$

$$\Pr(k \mid \mathbf{x}_\ell, \lambda_i) = \frac{w_i^{(k)} \mathcal{N}(\mathbf{x}_\ell; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)})}{\sum_{n=1}^{K} w_i^{(n)} \mathcal{N}(\mathbf{x}_\ell; \boldsymbol{\mu}_i^{(n)}, \boldsymbol{\Sigma}_i^{(n)})} \qquad (5)$$

where $\mathbf{x}_\ell$, $1 \le \ell \le L$, are the MFCCs of the available adaptation (singing) data; $w_i^{(k)}, \boldsymbol{\mu}_i^{(k)}$, and $\boldsymbol{\Sigma}_i^{(k)}$ are the weight, mean vector, and covariance matrix of the *k*-th mixture of λ_i, respectively; $\hat{\boldsymbol{\mu}}_i^{(k)}$ is the resulting mean vector after the adaptation; $\mathcal{N}(\cdot)$ is a multivariate Gaussian density function; γ is a weighting factor of the *a priori* knowledge to the adaptation data; and *K* is the number of mixture components. The improved system based on MAP adaptation of a speaker GMM to a singer GMM is shown in Fig. 2.

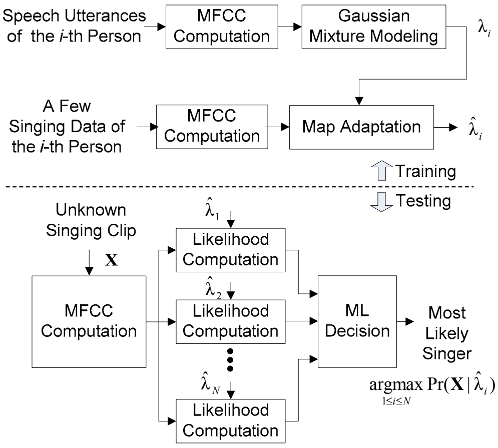

Fig. 2. The improved system based on MAP adaptation of a speaker GMM to a singer GMM.

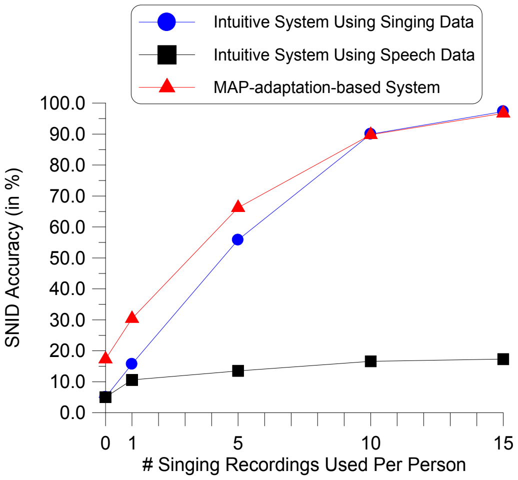

Fig. 3. SNID accuracies.

IV. Experiments

Because no public corpus of music recordings currently meets the specific criteria we set up for this study, a database of test recordings was created by ourselves. The database contains vocal recordings by twenty male amateur singers between the ages of 20 and 39. We asked each singer to perform 30 passages of Mandarin pop songs using a Karaoke machine in a quiet room. All the passages were recorded at 22.05 kHz, 16 bits, in mono PCM wave. The Karaoke accompaniments were output to a headset and were not captured in the recordings. The duration of each passage ranges from 17 to 26 seconds. We denoted the resulting 600 recordings by DB-Singing. Next, we asked each singer to read the lyrics of the 30 song passages at a normal speed. All the read utterances were recorded using the same conditions as those in DB-Singing. The resulting 600 utterances were denoted as DB-Speech. For use in different purposes, we divided DB-Singing into two subsets, DB-Singing-1 and DB-Singing-2, where the former contains the first 15 recordings per singer, and the latter contains the last 15 recordings per singer. Similarly, DB-Speech was divided into subsets DB-Speech-1 and DB-Speech-2, where the former contains the first 15 speech utterances per singer, and the latter contains the last 15 speech utterances per singer.

Fig. 3 shows the experiment results. Here, experiment of “Intuitive System Trained Using Singing Data” was conducted in the following way. We used the singing recordings in DB-Singing-1 to train the person-specific GMMs, and then tested each singing recording in DB-Singing-2. The number of Gaussian components used in each GMM was tuned to optimum according to the amount of the training data. To obtain a statistically-significant experiment result, we repeated the experiment by using the singing recordings in DB-Singing-2 to train the person-specific GMMs, and then tested each singing recording in DB-Singing-1. Thus, there were a total of 600 trials (20 singers × 15 recordings × 2 folds). The SNID performance was assessed with the accuracy:

$$\text{SNID Accuracy (in \%)} = \frac{\#\,\text{correctly - identified recordings}}{\#\,\text{testing recordings}} \times 100\% .$$

As to the experiment of “Intuitive System Trained Using Speech Data”, we used the speech utterances in DB-Speech-1 to train the person-specific GMMs, and then tested the singing recordings in DB-Singing-2. Also, the experiment was repeated by using the speech utterances in DB-Speech-2 to train the person-specific GMMs, and then testing the singing recordings in DB-Singing-1. There were a total of 600 trials.

In the experiment of “MAP-adaptation-based System”, we used the 15 speech utterances per person in DB-Speech-1 to train the person-specific GMMs. Each GMM was then adapted using J randomly-selected singing recordings per person in DB-Singing-1, where J = 1, 5, 10, and 15. Based on the adapted GMMs, the system identified the singing recordings in DB-Singing-2. In addition, to obtain a statistically-significant experiment results, we repeated the experiment by using DB-Speech-2 as the training data, DB-Singing-2 as the adaptation data, and DB-Singing-1 as the testing data. The identification accuracy was then computed as the percentage of the correctly-identified recordings.

It can be seen from Fig. 3 that, as expected, the more the training data, the better the performance is. We can also see that “Intuitive System Trained Using Speech Data” performs rather poor in identifying singing recordings. This is mainly because a person's voice can be different significantly between speaking and singing. Fig. 3 also shows that the “Intuitive System Trained Using Singing Data” performs very well when the amount of singing data for training is sufficient, but the system’s performance deteriorates largely as the amount of singing data decreases. Compared with the above two cases, “MAP-adaptation-based System” performs significantly better than both the intuitive system trained using speech utterances and the intuitive system trained using very limited singing data. Here, the case of 0 singing recording used in the MAP-adaptation-based system represents the result of the intuitive system trained using 15 speech utterances per person. It can be seen that the MAP-adaptation-based system improves the performance of the intuitive system trained using speech utterances, as long as singing data are available from each singer. The MAP-adaptation-based system also performs better than the intuitive system trained using insufficient singing data, i.e., the case of 1 and 5 recordings. This result validates our idea on singer voice characterization based on spoken data for SNID.

V. Conclusion

This study has investigated the feasibility of characterizing singers’ voices using their spoken data for SNID. Our experiment found that a GMM derived from a person’s speech utterances usually cannot well characterize his/her singing voice, owing to the significant difference between singing and speech. To overcome this problem, we have proposed a MAP-adaptation-based method to bridge the difference, so that a test singing recording can be identified using the speech-derived singer models. Although the proposed solution pays the cost for acquisition of a few singing data, its performance is significantly better than that of the system trained directly using such a few singing data.

References

[1] J. P. Campbell, “Speaker recognition: a tutorial,” in *Proc. IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.

[2] W. H. Tsai and H. M. Wang, “Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals,” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 14, no. 1, pp. 333–341, 2006.

[3] T. L. Nwe and H. Li, “Exploring vibrato-motivated acoustic features for singer identification,” *IEEE Trans. Audio, Speech*, Lang. Process., vol. 15, no.2, pp. 519-530, 2007.

[4] D. Gerhard: “Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing,” *J. Canadian Acoust. Soc*, vol. 30, no. 3, pp. 152-153, 2002.

[5] D. Gerhard, “Computationally measurable differences between speech and song,” Ph.D. dissertation, Simon Fraser University, 2003.

[6] J. Bonada and X. Serra, “Synthesis of the singing voice by performance sampling and spectral models,” *IEEE Signal Process*. Mag, vol. 24, no. 2, pp. 67–79, 2007.

[7] T. Saitou, M. Unoki, and M. Akagi, “Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis,” *Speech Comm*, vol. 46, pp. 405–417, 2005.

[8] T. Saitou, M. Goto, M. Unoki, and M. Akagi, “Speech-to-singing synthesis: vocal conversion from speaking voices to singing voices by controlling acoustic features unique to singing voices,” in *Proc. WASPAA*, pp. 215–218, 2007.

[9] D. Reynolds and R. Rose speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech Audio Process*, vol. 3, no. 1, pp. 72-83, 1995.

[10] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Statist. Soc*, vol. 39, pp. 1–38, 1977.

[11] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process*, vol. 10, pp. 19-41, 2000.

[12] *Robust text-independent*